A Personalized Support Agent for Depressed Patients

Kop, R.; Hoogendoorn, M.; Klein, M.C.A.

published in
Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 03
2014

document version
Publisher's PDF, also known as Version of record

citation for published version (APA)
Kop, R., Hoogendoorn, M., & Klein, M. C. A. (2014). A Personalized Support Agent for Depressed Patients. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 03* (pp. 302-309). IEEE Computer Society.

A Personalized Support Agent for Depressed Patients

Forecasting Patient Behavior using a Mood and Coping Model

Reinier Kop, Mark Hoogendoorn, and Michel C.A. Klein
VU University Amsterdam, Department of Computer Science
De Boelelaan 1081, 1081 HV Amsterdam, the Netherlands
r.kop@vu.nl, m.hoogendoorn@vu.nl, michel.klein@vu.nl

***Abstract*— Depression is a disorder that has a huge impact on both the patient and its environment. An effective treatment of depression is of crucial importance. Currently, Internet-based self-help therapies are the state-of-the-art among therapies that do not involve a human therapist. However, these interventions are not tailored towards individual patient needs. The utilization of pervasive technology, including a mobile phone and its sensors could potentially provide a way to make therapies more personalized and accessible at any time. One crucial aspect to make such personalization possible is to understand the current state of the patient and the ability to make a prediction on the expected state of the patient in the future. Obviously, predictions can differ greatly per patient. This paper takes a cognitive modeling approach in which the parameters of the model can be adapted to the characteristics of the patient. Hereby, an existing model for mood and coping is taken as a basis and different techniques are proposed to tailor the model towards the patient using sensory information that has been obtained. An evaluation is performed using a dataset from the psychological domain.**

Keywords— depression, prediction, mood and coping model

I. INTRODUCTION

Depression is a mental disorder associated with huge losses of quality of life in patients and their relatives, including increased mortality rates, high levels of service use, and enormous economic costs. Hence, effective treatment of depressive disorders is of utmost importance. Nowadays, a movement can be seen that is directed towards more self-help therapies within the domain of depression (see e.g. [15]). Such therapies facilitate the patient to schedule his/her own therapy, do homework assignments and obtain feedback without necessarily involving a human therapist in the loop. It has been shown that such treatments are as effective as face-to-face counseling [1]. Although clearly a promising result, the therapies are setup in a generic way, without much personalization towards the patient. A sophisticated mobile application which can be accessed by the patient at any time and utilizes sensory information to provide highly personalized and situation specific feedback and exercises might be able to bring a substantial added value and could truly engage the patient in the therapy.

In order for the mobile application to be effective, it should be able to build up a picture of the depressed patient, and also have the ability to forecast their expected developments. Based on this information it is possible to adapt the therapy and provide suitable feedback to maximize the effectiveness of the treatment. The composition of such a picture is however not a trivial task; each patient has his/her own characteristics and behaves in a different way. The approach chosen here is to use a cognitive model to make a judgment of the current and future state of the patient. This model has been developed in previous work [4; 5] and contains the most important states that play a role in depression and the relationship between these states. An experiment with six patients has shown that good parameter values could be found to describe the patient behavior well and even enable forecasting of the developments of the patient [3]. Although these results were promising, the learning techniques deployed were relatively simple, and only two states out of a total of ten were actually measured and forecasted. Furthermore, only a limited number of patients were used. To really understand the full capabilities and suitability of the model a more extensive evaluation is needed as well as a more sophisticated learning approach which potentially also utilizes a knowledge about a set of similar patients to tailor the model towards a new patient.

In order to tackle the issues as expressed above, there is a need for a richer, more extensive, dataset describing the behavior of depressed patients. Luckily, an increasing number of experiments are being conducted by researchers in the domain of Psychology in which so-called ecological momentary assessments (EMA) are frequently performed using mobile phones with a questionnaire application [14]. In the future, such a questionnaire could be partially replaced with advanced sensors which do not require explicit action of the patient. Even the current mobile applications however already facilitate the measurement of the key states of the patient on a regular basis. Such a dataset has been used in this paper, which allows for: (1) judging the descriptive and predictive capabilities of the existing model for a large number of different patients, and (2) developing a new algorithm that facilitates the personalization of a predictive model for mood and coping in a more effective way by exploiting data about historical patients. In the end, the model with the newly proposed

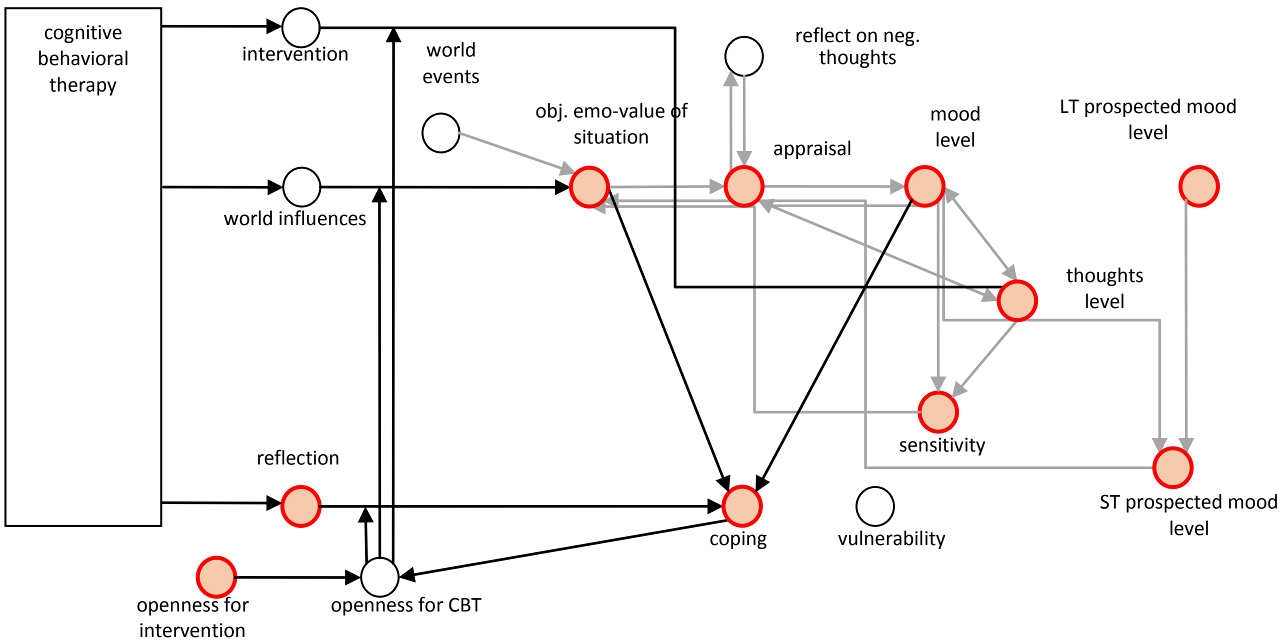

Fig. 1. Computational model for Cognitive Behavioral therapy.

learning techniques can form the true engine of a mobile computing application to support depressed patients.

This paper is organized as follows. Section 2 describes the mood and coping model. In Section 3 the dataset that has been obtained is described in more detail, whereas Section 4 presents the learning algorithms. Section 5 presents the results and finally, Section 6 concludes the paper.

II. MOOD AND COPING MODEL

In this paper, the same model for mood is used as in [3], which is based on [4] and [5]. A detailed description of the model can be found in those papers.

The main concepts include the *mood level*, *appraisal* and *coping skills* of a person, and how the levels for these states influence the external behavior in the form of selection of a situation. The model is based upon a number of Psychological theories, see [4] for a mapping between the literature and the model itself. In addition to the internal concepts that relate to emotions, a number of external influences are also part of the model: some things might not be controllable for the human (e.g. a negative event such as a sudden illness of a close family member). In this model, these external influences have been added specifically for a type of depression therapy, namely cognitive behavioral therapy (since this is the case study from which the data originates). However, they could just be seen as external influences on the global level.

In the model, a number of states are defined, where each state is represented by a number on the interval [*0,1*]. First, the state *objective emotional value of situation* is present, which represents the value of the situation a human is in (without any influence of the current state of mind of the human). The state *appraisal* represents the current judgment of the situation given the current state of mind (e.g. when you are feeling down, a pleasant situation might no longer be considered pleasant). The *mood level* represents the current mood of the human, whereas *thoughts level* the current level of thoughts (i.e. the positivism of the thoughts). The *long term prospected mood level* expresses what mood level the human is striving for in the long term, whereas the *short term prospected mood level* represents the goal for mood on the shorter term (in case you are feeling very bad, your short term goal will not be to feel excellent immediately, but to feel somewhat better). The *sensitivity* indicates the ability to select situations in order to bring the *mood level* to the *short term prospected mood level*. *Coping* expresses the ability of a human to deal with negative moods and situations, whereas *vulnerability* expresses how vulnerable the human is for negative events and how much impact that structurally has on the mood level. Finally, *world event* indicates an external situation which is imposed on the human (e.g., losing your job).

In addition to the states mentioned above, a number of states have been added to the model that are specific for a therapy that tries to influence the mood of the human [5]. First, a state representing the intervention or therapy (i.e., *intervention*) expresses that an intervention is taking place. The state *reflection on negative thoughts* expresses the therapeutic effect that the human is made aware of negative thinking about situations whereas the *appraisal effect* models the immediate effect on the appraisal of the situation. The *world influences* state is used to represent the impact of a therapy aiming to improve the *objective emotional value of situation*. The *openness for intervention* is a state indicating how open the human is for therapy in general, which is made more specific for each specific influence of the therapy in the state *openness for AS* (where AS stands for Activity Scheduling). Finally, *reflection* represents the ability to reflect on the relationships between various states, and as a result learn something for the future.

The states as explained above are causally related, as indicated by the arrows in Figure 1. These influences have been mathematically modeled (cf. [4; 5]).

III. EXPERIMENTAL DATA

The dataset used in this paper consists of 109 depressed patients, each of them part of either a control group (no treatment), an active control group (learning problem solving strategies and introspection) or undergoing emotion regulation training, based on cognitive behavioral therapy. For about a week long, each patient filled in measures of agreement to an extensive list of 122 statements, such as "I feel tired or without energy". The patients did this on a regular basis, so each day around ten answers to each of the questions were available for each patient. This collection of data is called the *pre-intervention data*. After this week, an intervention took place. Roughly six weeks later, the patients again answered the same statements for about a week long, the *post-intervention data*.

To use these statements, a mapping has been created between them and the concepts in the computational model. A mapping states that a specific statement in the survey was relevant for a concept in the model. Most of the concepts could be mapped to at least one survey statement, but because the data was not tailored towards this model, it was impossible to relate all model concepts to statements in the survey. As an example, the *objective emotional value of situation* could not be mapped, since the survey did not include statements regarding the situation the person found him- or herself in. As a consequence, it was not possible to derive a value for this concept from the survey data. Concepts with similar issues were *long term prospected mood level* and *openness to intervention*.

In the dataset, the measure of agreement is represented as a scale of either four or five possible answers. To derive values for the concept in the model, the data was first transformed to a value in the range [*0,1*]. Since all concepts depend on multiple statements in the survey, a mechanism was needed to aggregate the survey values that are relevant for a specific concept. The following algorithm has been used:

```
current_concept_value = 0.5
for each (survey_value relevant_for concept):
        if (current_concept_value – survey_value) ≥ 0:
                        new_concept_value = new_concept_value +
                        (1-current_concept_value) *
                        (survey_value-current_concept_value)
        else: new_concept_value = new_concept_value +
                        current_concept_value *
                        (survey_value-current_concept_value)
        end if
end for
```

The algorithm expresses that the default value of each concept is 0.5. The *new concept value* will increase (to a maximum of 1) compared to the *current concept value* in case the encountered *new survey value* is higher than the *current concept value*, and decrease (to a minimum of 0) in the opposite case. These calculations result in a value for the states of the model (except for the concepts specified earlier) for each time point at which the full questionnaire was filled in by the patient, i.e. a series of values for the first week and a series of values for the last week.

After this pre-processing, a running average was taken over the various time points, to smoothen the course of the actual values. This approach simultaneously solved the problem of any missing values the answers contained.

A total of nine concepts were derivable from the dataset using this approach, seven of which eventually suitable for calculating performance (see section IV).

IV. LEARNING ALGORITHMS

Given that the main goal is to personalize the mobile application using the model, the parameters of the model should be tailored towards the specific user. The aim of the learning algorithms is to find sets of parameters that results in a model which provide maximal predictive capabilities. These predictions form the basis for dedicated feedback and an engaging user experience. For the learning algorithms, the following hypotheses are expressed:

H1. Utilizing data from other patients can help in improving performance of the prediction for individual patients.

H2. Individualizing parameter values (as opposed to having a single set of parameters for all patients) will result in better performance with respect to the *reproduction* of patterns seen for a patient.

H3. Individualizing parameter values will result in an improved performance with respect to the *prediction* of future patterns.

Given these hypotheses, different variants of learning algorithms have been specified. The basis for the learning is formed by a Genetic Algorithm (GA). The precise application of the Genetic Algorithm to the case at hand is expressed first, followed by different variants of the learning algorithm.

A. Learning Basis

The genetic algorithm applied is the Java GA Package, or JGAP[1]. The main features of a GA are discussed below.

- Chromosome representation: each individual within the population is represented by a vector pv: <pv_1, …, pv_n>, where pv_i is the parameter value of the i-th parameter used by the mood and coping model. Initialization is done by taking random real values between 0 and 1.
- Selection: after evolving, the X% individuals (both parents and children) with the highest fitness are chosen. The 100-X% left are clones of the top performing individuals.
- The crossover operator is a simple one-point crossover, requiring two parents and creating two children in the process.
- Mutation: each gene is mutated with a probability *p*.

In order to evaluate the effectiveness of a chromosome in the population (i.e. assign a fitness value, in GA terms), the parameters are used to simulate the entire model (given certain initial values for the states). This results in a temporal pattern of prediction values between the simulation start time t_{start} and the simulation end time t_{end} for each of the individual states within the model: pred_value(state, t, pv). For example if the state *mood* has a predicted value of 0.5 at t_1 with parameter vector pv, then pred_value(mood, t_1, pv) = 0.5. In addition, it is assumed that there is a certain observed

[1] http://jgap.sourceforge.net/

TABLE I : LEARNING ALGORITHMS. THE PRE-INTERVENTION WEEK IS WRITTEN AS PRE-WEEK. THE POST-INTERVENTION WEEK IS WRITTEN AS POST-WEEK.

Approach	**Training set**	**Test set**	**Test patient parameter selection**
Initial learning period	Pre-week, 100% of patients	Post-week, 100% of patients	Same patient
Generic parameters	Pre- and post-week, X% of patients	Pre- and/or Post-week, 100-X% of patients	Global parameter set
Nearest Neighbor	Pre- and post-week, X% of patients	Pre- and/or Post-week, 100-X% of patients	Same as most similar (= nearest neighbor)

value for each state (coupled to an individual patient p): obs_value(state, t, p). The fitness value is then calculated by first taking the root mean squared error (RMSE) as follows:

$$RMSE(pv, p) = \sqrt{\frac{\sum_{t=t_{start}}^{t_{end}} \sum_{\forall s:state} (obs\_value(s,t,pv) - pred\_value(s,t,p))^2}{\#timepoints \cdot \#states}}$$

This value is then converted to a fitness value (where a higher value means a better set of parameters) by simply applying the following formula:

$$Fitness(pv, p) = (1 - RMSE(pv, p)) * 100$$

The fitness is thus essentially equal to the accuracy of certain parameter sets in percentages.

B. *Learning Algorithms*

Given the specification above, learning can now be applied to find parameter sets with good predictive capabilities. A number of different approaches to optimize the parameters for an individual patient are specified in this paper. The following variants of learning are distinguished:

1. *Initial individual learning period* – Learn the parameters at an individual level during a learning phase, and use those to predict future states. In this case, the learning phase (i.e. the training set) is the pre-intervention survey week for the specific patient and the fitness function to be optimized is based on the RMSE of that training set. The testing phase (i.e. test set) is the post-intervention survey week. This approach is similar to [3] and is mainly investigated to continue the exploration of this model in that respect. Note that in this approach, no information is shared regarding parameter settings based on other patients which may be similar. For each new patient, optimal parameter settings need to be found. Previously learned data is thus not incorporated in the prediction, which gives at the very least a cold start problem when patients start using for example a mobile application which incorporates an approach such as this one.
2. *Generic Parameters* – This approach is drastically different from approach 1. First, only a single parameter set is learned for all patients. Second, the learning phase here is both the pre- and post-intervention week of all patients in the *training* set. The fitness function is extended to be the combined RMSE across all patients in the training set. The patients in the *test* set then use the parameter set obtained in the training set, after which the test accuracy is calculated. In summary, one set of parameters that works well across all patients in a certain training set will be used for predictions of states in a test set.
3. *Nearest Neighbor Parameters* - To be able to include information about specific patients that have been seen before in the tuning of the parameters, a third approach is proposed which takes the known parameters of another patient that is closest to the current patient. This approach is very similar to approach 2 in how the model is trained. Thus, the learning phase here is again both the pre- and post-intervention week of all patients in the *training* set. However, an important difference is the fact that, like in approach 1, the model finds parameter sets *for each individual patient*. The patients in the test set are then assigned parameter values based on similarity between themselves and patients in the training set. For each of the patients in the test set, the full parameter set of the most *similar* patient in the training set is chosen, and the test accuracy is calculated. This *similarity* between a patient in the test set and one in the training set is calculated based on abstracted values from the pre-intervention week only. This is done so the approach still has proper predictive capabilities. For each concept in this week, its mean, variance, and the amount of change (which is simply the mean of the first half of the pre-intervention week subtracted from the mean of the second half) will be calculated.

Table I shows the characteristics of the learning approaches and the corresponding settings of the patients, where pre- and post-week represent the pre- and post-intervention data, respectively.

V. RESULTS

A. *Algorithm parameters*

The genetic algorithm and nearest neighbor approach described in the previous session require some parameters to be set, themselves. For the genetic algorithm, the default configuration provided by JGAP is used, which accumulates to a random point crossover with a rate of 0.35, a mutation rate of 0.08 (per gene) and a 90% elitist selection. These parameter settings were found to be best after a substantial period of trying various parameter settings for the GA. Furthermore, each chromosome contains 15 double values (equal to the amount of parameters we are tuning) between 0 and 1. A simple 1-nearest neighbor approach was used, because the GA provided such diverse sets of parameters that averaging parameters across multiple neighbors did not seem promising.

TABLE II : MEAN ACCURACIES AND STANDARD DEVIATIONS FOR ALL APPROACHES.

Approach		$\mu_{training}$ (%)	$\sigma_{training}$ (%)	μ_{test} (%)	$\sigma_{training}$ (%)
1. Initial learning period		88.91	4.46	72.35	10.93
2. Generic parameters	Pre-week only	84.00	5.66	84.11	5.55
	Post-week only	79.08	6.89	79.36	6.71
	Aggregated	81.54	6.76	81.74	6.56
3. Individualized Parameters	Pre-week only	87.54	4.29	81.47	7.00
	Post-week only	86.13	4.46	74.42	8.61
	Aggregated	86.84	4.43	77.95	8.56

TABLE III : COMPARISON BETWEEN APPROACH 1 ON THE ONE HAND, AND APPROACHES 2 AND 3 ON THE OTHER HAND

Approach Comparison	**p_value < α?**	**Best performing**
(1) vs (2), test, post-week only	yes (p = 0.0012)	Approach 2
(1) vs (3), test, post-week only	no (p = 0.3607)	No difference

TABLE IV : DESCRIPTIVE (TRAINING) PHASE COMPARISON BETWEEN APPROACH 2 APPROACH 3

Approach Comparison	**p_value < α?**	**Best performing**
(2) vs (3), training, post-week only	yes (p < 0.0001)	Approach 3
(2) vs (3), training, aggregated	yes (p < 0.0001)	Approach 3

TABLE V : PREDICTIVE (TEST) PHASE COMPARISON BETWEEN APPROACH 2 APPROACH 3

Approach Comparison	**p_value < α?**	**Best performing**
(2) vs (3), test, post-week only	yes (p = 0.0067)	Approach 2
(2) vs (3), test, aggregated	yes (p = 0.0336)	Approach 2

B. Dataset

Since the model is constructed to be for cognitive behavioral therapy, only the self-regulation training group (a form of CBT) is investigated. Thirty patients in this set were removed because of erroneous dates (e.g. 1-1-1970 for some or all entries) or a lack of occurrence in either the pre- or post-intervention week, or both. This left us with 38 patients to experiment with. 10-fold cross validation was used for approaches (2) and (3), whereas for approach (1), simply all the pre-intervention week data is the training data, and all the post-intervention week data is the test data (no cross fold validation is possible in this case).

C. Accuracies

The resulting mean training and test accuracies and corresponding standard deviations are summarized in table II. The different rows for the second and third approaches show accuracies of the individual parts of the data set. The main interest of this paper is in predicting the post-intervention week. The aggregated accuracy incorporates all data set entries, both from the pre- and the post-intervention weeks. All statistical comparisons described below were made using unpaired two-tailed t-tests with a statistical significance level of α=0.05.

D. Validation of hypotheses

Hypothesis 1 is concerned with whether approaches 2 and/or 3 perform better than approach 1. We may compare these approaches because the test set of approach 1 is identical to the test set of approaches 2 and 3 when only regarding the latter's post-week accuracies. In table III, we see that the data partly support hypothesis 1. Namely, approach 1's testing phase (μ=72.35%, σ=10.93%) is significantly smaller than approach 2's post-intervention testing week (μ=79.36%, σ=6.71%). However, approach 1's testing phase is statistically the same as approach 3's post-intervention testing week (μ=74.42%, σ=8.61%). It can be concluded that, depending on the exact approach taken, incorporating data from previous cases *can* help in predicting new cases.

Hypothesis 2 is concerned with whether the individualization of parameters gives en edge to the *descriptive* performance (i.e. *training* set performance) of the model. In order to investigate this, we compare the performance of the training sets of both the post-intervention week and the aggregated results for approaches 2 and 3. In table IV, it can be seen that the data support hypothesis 2: approach 2's training phase for the post-intervention week (μ=79.08%, σ=6.89%) is smaller than approach 3's training phase for the post-intervention week (μ=86.13%, σ=4.46%). This is also the case for the aggregated accuracy (μ=81.54%, σ=6.76% is smaller than μ=86.84%, σ=4.43%).

Hypothesis 3 is identical to hypothesis 2, except for the fact it is concerned with the *predictive* capabilities (or *test* set performance) of the model. Thus, the post-intervention week and aggregated performances of the test sets of approaches 2 and 3 are compared. In table V, it can be seen that the data do not support hypothesis 3. Namely, approach 2's test phase for the post-intervention week (μ=79.36%, σ=6.71%) is larger than approach 3's test phase for the post-intervention week (μ=74.42%, σ=8.61%). This is also the case for the aggregated accuracy (μ=81.74%, σ=6.56% is larger than μ=77.95%, σ=8.56%).

E. Analysis

In order to get a better understanding of the results, the patients for which the performance is best have been selected. Figure 1 represents the actual level of the state *thoughts* as shown by the patient versus the predicted level of *thoughts* using the initial learning approach (approach 1). Notice that the initial period used for learning includes the set of points on the left, whereas the set of points on the right are the ones that have not been seen before and are predicted given the initial period. Figure 2 shows the same setup for the *appraisal* state. The figures show that the initial approach seems to predict the trend in quite a reasonable way without doing it on a very detailed level. However, the prediction is quite difficult as there is a substantial period of missing data and the first period is not necessarily representative nor a good predictor for the second period.

Figures 3 and 4 show examples of approach 2 (using one generic set of parameters) for the *mood* level and *sensitivity* respectively. The graphs show that a straight line is specified by the model which can be explained by the fact that the average difference between the pre and the post-week is almost zero when taking a large group of patients. Hence, predicting the same level in the pre- and post-week makes sense.

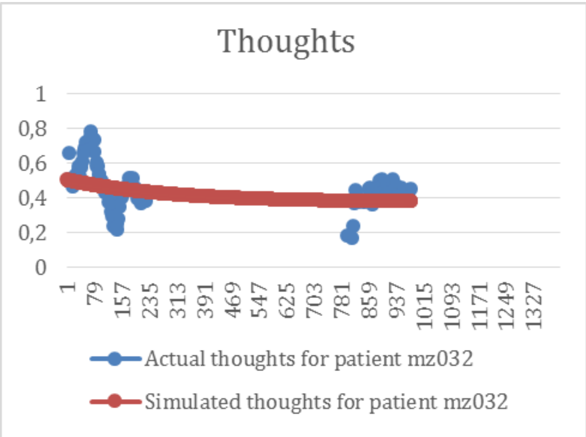

Fig. 1. Actual and predicted thoughts level for patient mz032 as calculated by approach 1 (initial learning period)

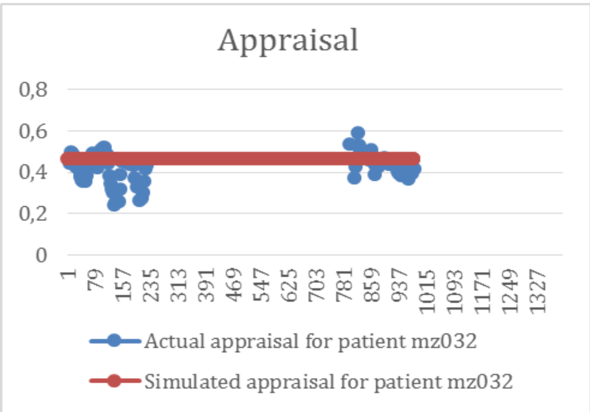

Fig. 2. Actual and predicted appraisal for patient mz032 as calculated by approach 1 (initial learning period).

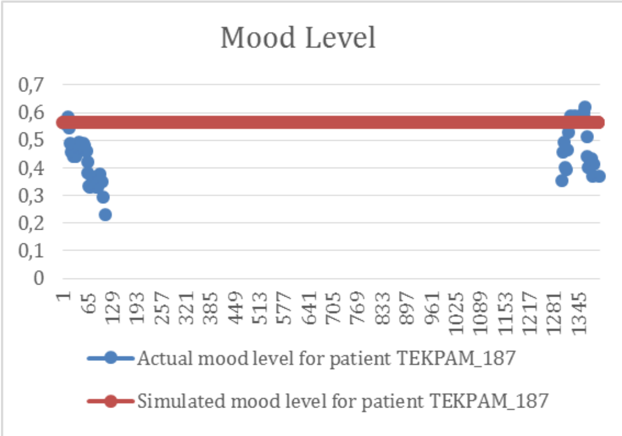

Fig. 3. Actual and predicted mood level for patient TEKPAM_187 as calculated by approach 2(generic parameters).

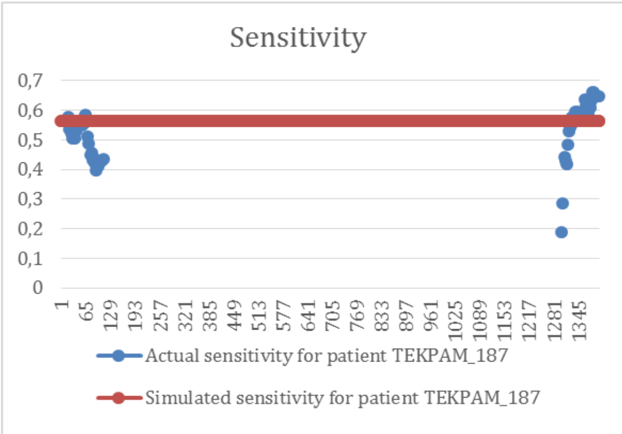

Figure 4 Actual and predicted sensitivity for patient TEKPAM_187 as calculated by approach 2 (generic parameters).

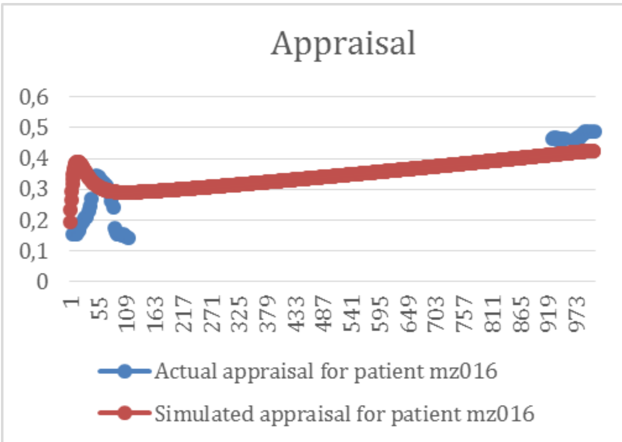

Fig. 5. Actual and predicted appraisal for patient mz016 as calculated by approach 3 (nearest neighbor).

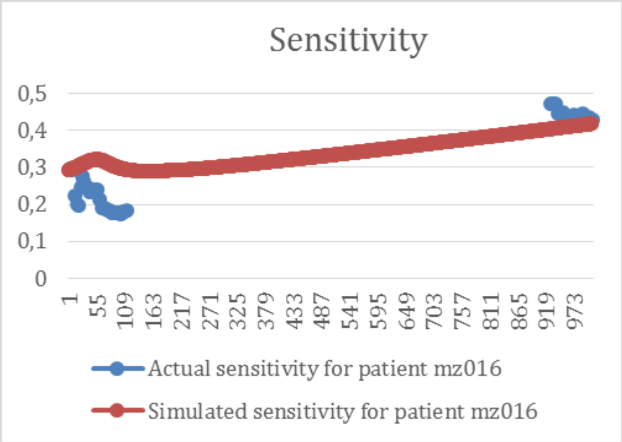

Fig. 6. Actual and predicted sensitivity for patient mz016 as calculated by approach 3 (nearest neighbor).

Finally, Figures 5 and 6 show the results for the nearest neighbor approach (approach 3), for the *appraisal* and *sensitivity* states, respectively. Here it can be seen that the trend is nicely predicted. When comparing these figures to figures 3 and 4, it may seem puzzling that according to table V, approach 2 outperforms approach 3. Two reasons could underlie this. First, the behavior shown in Figure 5 and 6 only holds for a limited number of patients, as the period to compare the values of a new patient with a known patient (using the 1-nearest neighbor algorithm) is very short (one week long, the pre-intervention week) and not necessarily representative for the values seen for the states in the later period. Second, the dataset may show too little variation in general between the pre- and post-intervention weeks. This might allow for a static straight line (as in figures 3 and 4) to provide reasonably good prediction, while in reality there is (subtle) change in the reported values of the patient.

VI. DISCUSSION AND CONCLUSION

Correctly predicting the course of a depression based on measurements of a patient's behavior and self-reports can be helpful to create a highly personalized support system which truly engages the patient. In order to make such predictions, in this paper a model-based approach has been taken as basis, which still requires tailoring towards the individual patient.

Only few computational models of depression have been developed [4; 6], and those have not been personalized based on actual patient data. A model of depression can be seen as a specific instantiation of a model of mood and emotion regulation. There are more computational models for emotions, for example [7; 9; 16]. However, also for these models holds that there has been no attempt to personalize them for individual patients based on real data.

There are a limited number of studies in which such models have been related to data about actual humans. Gratch and Marsella [10] utilized the aggregated outcomes of a questionnaire with respect to stress and coping to evaluate their proposed model, but not to adapt it. In [11] the relation between appraisals and the intensity of human emotion is measured, and compared with several emotion models. In both papers, the main intent is not to look at individuals, but to look at general trends of the emotions of humans based on the situation (although in [11], subgroups were studied).

Another body of research involves the evaluation of *systems that are using a model* of emotion or mood. Most of these evaluations take place based upon user experience and not on the realism of the emotional levels themselves (e.g. measuring actual mood or emotion levels in users). Examples of papers describing such an evaluation are for instance [2], [8] and [13]. Pontier & Siddiqui [12] performed such an evaluation specific for a virtual therapist incorporating emotions in depression. Their results showed that the emotion-based virtual agent was preferred over the non-emotional agent.

In this paper, however, a more ambitious approach was taken. The aim was to fully tailor the parameters of the model towards the behavior of the human, possibly incorporating historical data from other patients, and showing that the model is able to describe the behavior of an arbitrary user with high accuracy. Especially when the purpose of the model is to make predictions about a specific individual, such an evaluation is crucial. Hereto, three techniques have been examined in this paper: (1) learning parameters for individual patients during an initial period in the data set; (2) finding parameters that work well across a lot of patients by learning during the complete period of the dataset, and (3) using parameters that worked well for prior similar patients.

By experimenting with an extensive dataset it was shown that using *more* information than just the initial period for finding appropriate parameters (i.e. approach 2 and 3) is better than using solely the initial information from a single patient (approach 1), as was conjectured in hypothesis 1. Do note however that for this dataset only generic parameters were shown to perform significantly better than those found by the initial learning period approach which was not anticipated in advance. The reason could be the highly complex and varying relationship between the pre-intervention week data and the data from the post-intervention week.

By comparing the predictions of a personalized model (approach 3) with model predictions based on a standard parameter set (approach 2), it has been shown that personalization can be beneficial for *descriptive* purposes (i.e. the results for the training sets), as was surmised in hypothesis 2.

However, for this specific dataset and model, a global parameter approach (approach 2) seemed to work better for making *predictions* than the nearest neighbor approach (approach 3), unlike what hypothesis 3 stated. Nevertheless, due to the fact that approach 3 provides more 'natural' values than approach 2 for some patients (when looking at the graphs), it is suggested that hypothesis 3 needs to be investigated more thoroughly on different data sets.

There are two possible explanations that the predictions for approach 3 were not as accurate as initially anticipated. First, there exists the possibility that the abstracted values for the pre-intervention survey week are not representative enough for a patient's (complex) behavior. Therefore, the nearest neighbor measure may not have been able to select the 'actual' most similar patient. Second, the nearest neighbor approach might have been too naïve after all, in that only a single neighbor was used. This was initially done because we assumed taking averages of *k* other patients (*k*-

NN) would create too much noise, since the parameters tuned were so vastly different from each other.

Apart from the possible imperfections mentioned in the previous paragraph, it might be interesting to look at giving different weights to states when calculating the fitness. The *mood* state could for instance be made more important for the fitness, which might change the final parameters chosen, affecting the performance (for better or worse).

ACKNOWLEDGMENTS

This research has been performed in the context of the EU FP7 project E-COMPARED (project number 603098). We would like to thank the Academy Assistants project of the Network Institute at the VU University Amsterdam, the Netherlands for providing the funding of this subproject. Furthermore, we would like to thank Professor Berking and Anna Radkovsky from the University of Marburg for sharing the dataset.

REFERENCES

[1] Andersson, G., Bergstrom, J., Hollandare, F., Carlbring, P., Kaldo, V., Ekselius, L. (2005). Internet-based self-help for depression: randomised controlled trial. British Journal of Psychiatry, 187, 456-61.

[2] Bosse, T., Zwanenburg, E.: There's Always Hope: Enhancing Agent Believability through Expectation-Based Emotions. In: Pantic, M., Nijholt, A., Cohn, J. (eds.), Proceedings of ACII'09, pp. 111--118, IEEE Computer Society Press (2009)

[3] Both, F., Hoogendoorn, M., and Klein, M.C.A., Validation of a Model for Coping and Mood for Virtual Agents. In: Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE Computer Society Press, 2012, pp. 382-389.

[4] Both, F., Hoogendoorn, M., Klein, M.A., and Treur, J., Formalizing Dynamics of Mood and Depression. In: M. Ghallab, C.D. Spyropoulos, N. Fakotakis and N. Avouris (eds.), Proceedings of the 18th European Conference on Artificial Intelligence, ECAI'08. IOS Press, 2008, pp. 266-270.

[5] Both, F., Hoogendoorn, M., Klein, M.C.A., and Treur, J., Computational Modeling and Analysis of Therapeutical Interventions for Depression. In: Yao, Y., Sun, R., Poggio, T., Liu, J., Zhong, N., and Huang, J. (eds.), Proceedings of the Second International Conference on Brain Informatics, BI'10. Lecture Notes in Artificial Intelligence, vol. 6334, Springer Verlag, 2010, pp. 274-287.

[6] Davidson, R.J., D.A. Lewis, L.B. Alloy, D.G. Amaral, G. Bush, J.D. Cohen, W.C. Drevets, M.J. Farah, J. Kagan, J.L. McClelland, S. Nolen-Hoeksema & B.S. Peterson (2002) Neural and behavioral substrates of mood and mood regulation. *Biological Psychiatry, Volume 52, Issue 6*, pp. 478-502.

[7] Dias, J., & Paiva, A. (2005). Feeling and reasoning: A computational model for emotional characters. In *Progress in artificial intelligence* (pp. 127-140). Springer Berlin Heidelberg.

[8] Gebhard, P., & Kipp, K. H. (2006). Are computer-generated emotions and moods plausible to humans?. In *Intelligent Virtual Agents* (pp. 343-356). Springer Berlin Heidelberg.

[9] Gratch, J., & Marsella, S. (2004). A domain-independent framework for modeling emotion. *Cognitive Systems Research*, *5*(4), 269-306.

[10] Gratch, J., and Marsella, S., Evaluating a Computational Model of Emotion. Autonomous Agents and Multi-Agent Systems, vol. 11, 2005, pp. 23-43.

[11] Gratch, J., Marsella, S., Wang, N., and Stankovic, B.. Assessing the validity of appraisal-based models of emotion. Int. Conf. on Affective Computing and Intelligent Interaction. A'dam, IEEE, 2009

[12] Pontier, M.A., and Siddiqui, G.F., A virtual therapist that responds empathically to your answers. In: Prendinger, H., Lester, J., and Ishizuka, M. (eds.), Proceedings of IVA'08, 2008, pp. 417-425.

[13] Pontier, M.A., Siddiqui, G.F., and Hoorn, J., Speed Dating with an Affective Virtual Agent - Developing a Testbed for Emotion Models In: J. Allbeck et al. (Eds.), Proceedings of IVA'10, 2010, Lecture Notes in Computer Science, Vol. 6356, pp. 91-103.

[14] Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, *4*, 1-32.

[15] Spek V, Cuijpers P, Nyklíček I, Riper H, Keyzer J, Pop V (2007). Internet-based cognitive behavior therapy for mood and anxiety disorders: a meta-analysis. Psychological Medicine, 37, 319-328.

[16] Velasquez, J.D. 1997. Modeling Emotions and Other Motivations in Synthetic Agents. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, 10-15. Menlo Park, Calif.: AAAI Press.